Journal of Biomedical Informatics 69 (2017) 1–9

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks

Shaodian Zhang [a,*], Edouard Grave [a], Elizabeth Sklar [b], Noémie Elhadad [a]

[a] Department of Biomedical Informatics, Columbia University, New York, NY, USA
[b] King's College London, London, UK

ARTICLE INFO

Article history:
Received 17 April 2016
Revised 14 March 2017
Accepted 16 March 2017
Available online 18 March 2017

Keywords:
Topic
Online health community
Breast cancer
Convolutional neural network
Deep learning
Longitudinal analysis

ABSTRACT

Identifying topics of discussions in online health communities (OHC) is critical to various information extraction applications, but can be difficult because topics of OHC content are usually heterogeneous and domain-dependent. In this paper, we provide a multi-class schema, an annotated dataset, and supervised classifiers based on convolutional neural network (CNN) and other models for the task of classifying discussion topics. We apply the CNN classifier to the most popular breast cancer online community, and carry out cross-sectional and longitudinal analyses to show topic distributions and topic dynamics throughout members' participation. Our experimental results suggest that CNN outperforms other classifiers in the task of topic classification and identify several patterns and trajectories. For example, although members discuss mainly disease-related topics, their interest may change through time and vary with their disease severities.

1. Introduction

The involvement of the Internet in healthcare gives rise to new perspectives in eHealth [1] and changes the way patients consume and contribute health-related information. Traditionally, patients with life-threatening conditions receive most of the information about their disease from their care providers. While providers tend to focus on the clinical impact of the disease and might ignore the impact of the disease on a patient's emotional wellbeing and daily life [2], support groups, and more recently online health communities (OHCs), can act as a complementary source of support for patients [3]. In particular, public online health communities such as Breast Cancer Forum [4–6], the CSN network [7,8], and Facebook groups [9] are increasingly popular among patients, and have produced an unprecedented amount of user-generated content which could be valuable resource for studying OHCs.

There are many challenges in understanding the very large amount of content authored and read by online health community members, however. Some relate to the quality of information, as well as how the information is consumed and integrated by community members into their daily lives and disease management decisions. One fundamental content-related task that is important to downstream content analysis is to identify topics of discussions [10]. Previous research suggested that topic, along with emotion, are two basic building blocks of content with respect to OHC [7]. In this study, we focus on investigating prevalence and dynamics of discussion topics in a popular online breast cancer forum. The task is challenging because topics discussed in such OHCs are usually heterogeneous and domain-dependent, and can be different from themes in other biomedical content such as clinical notes, as well as those in other types of general-purpose communities such as Facebook. Previously, topic classification has also been a central issue of text mining in general [11]. Few studies have focused on automated topic classification for online health communities [12,7,13], but to our best knowledge there was no previous research that addressed both the multi-label classification problem and the issue of temporal dynamics of member participation.

In this paper, our study objectives are (i) to provide an annotation schema for topic classification; (ii) to contribute an annotated dataset of sentences and posts according to the coding schema; (iii) to experiment with different supervised classification tools, including convolutional neural networks, support vector machines, and labeled latent Dirichlet allocation, to automate the annotation process; and (v) to explore the prevalence and dynamics of different discussion topics in the entire breast cancer community and across member with different disease severities. Specifically, we ask following research questions:

* Corresponding author.
E-mail addresses: sz2338@columbia.edu (S. Zhang), edouard.grave@gmail.com (E. Grave), elizabeth.sklar@kcl.ac.uk (E. Sklar), noemie.elhadad@columbia.edu (N. Elhadad).

http://dx.doi.org/10.1016/j.jbi.2017.03.012
1532-0464/

1. What is the most effective supervised learning tool in classifying topic of discussions in an online health community?
2. What are the most prevalent topics in discussions in the breast cancer forum?
3. Are there any differences of topic foci among patients of different cancer stages?
4. How does the distribution of topics change through time, as members participate longer in the community?

1.1. Related work

Previously, Sharf observed that in an online breast cancer group, topics regarding basic classifications or definitions of tumors and diagnosis are most prevalent, indicating that Internet support was primarily a complementary source of information in early years [14]. A variety of themes such as relationship/family issues became popular in online peer discussions according to subsequent studies conducted more recently [15,16], but disease specific topics like treatment, diagnosis, and interpretation of lab test results are still most prevalent [17–19]. Specific topics of discussion were identified as well. For example, based on content analysis, Meier and colleagues found that the most common topics in 10 cancer mailing lists were about treatment information and how to communicate with healthcare providers [18]. Owen and colleagues proposed a topic schema which includes seven categories: outcome of cancer treatment, disease status and processes associated with the cancer, healthcare facilities and personnel, medical test and procedures, cancer treatment, physical symptoms and side effects, and description of cancer in the body [16]. Based on such schema, prevalence of different topics can be quantified to facilitate content analysis of cancer support groups. More recently, relying on quantitative methods, topic modeling is carried out for public OHCs, but in an unsupervised fashion [7].

2. Methods

2.1. Source of data and data processing

Our work was approved by the Columbia University IRB office. We relied on the discussion board of the publicly available community from `breastcancer.org`. The entire content of the discussion board was collected in January 2015. The discussion board is organized in distinct forums, each with threads and posts. In total, the data set consists of 3,283,016 posts from 121,474 threads, authored by 58,177 members. The following pre-processing steps were carried out.

For each post, meta-data about the forum and the thread in which it was authored was kept, along with author and creation date. The content of each post was pre-processed by (i) removing all non-textual content (e.g., substituting emoticon icons with emoticon-related codes); and (ii) identifying sentence boundaries using the open-source tool NLTK [20].

In addition to the post content, we also collected signatures of posts (see Fig. 1 for an example), which consist of self-reported disease information of patients. This include diagnosis and treatment histories of members. Notice that not all members have such information available. In our analysis we will be using one specific disease variable, cancer stage. In total we successfully collected stage information for 7211 members.

2.2. Creating the topic schema

To enable reliable and useful annotation of topics, we established a coding schema of discussion topics through a literature review of information needs in online health communities, with an emphasis on breast cancer communities [18,17,21–25]. Our objectives were (i) to devise a coding scheme that is both relevant to describing the information needs of community members as well as applicable to and robust enough for automatic topic classification; and (ii) to design a coding scheme that can be applied to characterizing topics of discussion for either an entire post or its individual sentences. Furthermore, the annotation schema is such that each unit of annotation can be labeled according to one or more topics. For instance, a given post, and even a given sentence can simultaneously convey information about a treatment and the health system.

Fig. 1. An example member signature, including cancer diagnosis and treatment history.

The coding scheme was developed using an iterative process to reflect the main topics of discussion of post content [26]. Preliminary coding of 439 sentences (corresponding to 37 posts) provided the initial categories and guidelines for coding. Upon review and discussion, infrequently used categories were collapsed into larger concepts, and the 439 sentences were coded again to verify sufficient agreement between the two initial coders. The 439 sentences and their codes were used as training instances for the later coders, along with the coding guidelines.

Our final topical scheme contains 11 topics, as listed in Table 1. It is noteworthy that the topics focus on informational support, rather than emotional dimensions and range from clinical to daily matters.

We also learned from the preliminary coding that members may shift topic of discussion in a post, which reminded us that to achieve better granularity sentence-level coding would be necessary. As such, our manual annotation described below were carried out at sentence level rather than post-level.

2.3. Manual annotation

We selected a subset of posts (1008 posts consisting of 9016 sentences) from the original dataset described above. The posts were selected from the different forums, where each forum focuses on specific aspects of breast cancer management, such as diagnosis and treatment options, support through chemotherapy, nutrition, alternative treatments, and daily life. Posts were thus grouped in batches of 50 posts per manual annotation session.

Sentences were coded according to double annotation followed by an adjudication step from one dedicated adjudicator throughout the annotated dataset. Three coders were hired for the annotation, all female native English speakers with undergraduate degrees. To train for the annotations, coders practiced annotating the 439 sentences (37 posts) referred to above using the annotation guidelines. Inter-annotator agreement with gold-standard topic annotation was monitored throughout training, and training was terminated when a coder had achieved a 0.6 Kappa (agreement statistic) with the gold-standard annotation [27]. Note that given the large number of potential labels in the schema and the fact that each sen-

Table 1
Annotation schema for breast cancer forum text.

Topic	Abbreviation	Description
Alternative	ALTR	Alternative and integrative medicine
Daily	DAIL	Daily cancer-related experience
Diagnosis	DIAG	Diagnoses, measurements, and results of tests
Finding	FIND	Health finding, sign, symptom or side effect
Health systems	HSYS	Health systems patients interact with, including nurses, doctors, practices, hospitals, and insurance companies
Miscellaneous	MISC	Greetings, uninformative sentence, or any sentence, which does not fit under any other annotation label
Nutrition	NUTR	Nutrition
Personal	PERS	Personal information
Resources	RSRC	Link, pointer, or quote towards an external information resource
Test	TEST	Testing procedures (but not results of tests)
Treatment	TREA	Treatments, including procedures, medications and therapeutic devices

tence can be labeled according to multiple topics, this is a particularly stringent training constraint. Afterwards, each batch of posts was assigned two coders and was doubly annotated at the sentence level. Finally, the adjudicator went through all posts, resolved differences between coders and made final decisions over sentence topic labels.

2.4. Topic classification

Because a given sentence in a post can be described according to multiple topics (e.g., a sentence can be about a treatment, nutrition, and daily matters all at once), the task of automating the topic coding can be cast as a multi-label classification: for each sentence, there can be up to N labels, where N is the number of topics in the schema. This type of classification is more challenging than single-label classification, where one sentence can be described by only one label chosen from the N topics in a schema. Traditionally, there are two approaches for multi-label, multi-class classification: problem transformation methods and algorithm adaptation methods [28].

In this paper, we rely on three different supervised classifiers, a labeled LDA classifier [29], an SVM [30], and a convolutional neural network [31]. They represent three types of mainstream supervised learning frameworks: generative graphical models, discriminative max-margin linear classifiers, and neural networks. Within these three models, labeled LDA and neural networks are able to handle multi-label classification naturally since they allow multiple outputs. For the SVM, we consider N binary, single-label classifiers and aggregates the N outputs into one multi-label.

For the labeled LDA classifier, we rely on an self-implemented Gibbs sampler for labeled LDA, based on the open source LDA implementation [32].[1] The two hyper-parameters of the model, alpha and beta, are set as 0.1 and 0.5 experimentally according to a grid search. For SVM, we rely on the open source tool LibSVM [33]. We also did a naive grid search (logarithm grid search for c and gamma, specifically) for parameters to find the setting which yields best performance. In the result section we will report performance of SVM under following parameters: radial basis function kernel with $c = 100$ and gamma $= 10^{-3}$, and radial basis function kernel with $c = 10$ and gamma $= 10^{-2}$, for SVM and SVM-e respectively. Other parameters were set as default because we observed that in this task performance of the SVM were only slightly influenced by them.

[1] http://jgibblda.sourceforge.net.

The convolutional neural network we used follows [31], which has one hidden convolutional layer. First, the sequence of words is represented as a sequence of vector of dimension $D = 100$, by using a lookup table. The 100-dimensional vectors are concatenate as in [31] (our model is equivalent to the "single channel" model described in Section 2 of this paper). The word embeddings used in this lookup table were pre-trained, by using the word2vec algorithm, on the entire unannotated dataset from the same forum. Then we take the convolutions of this sequence of "word vectors" with H filters, obtaining a score for each filter and each position in the sentence. In order to obtain a fixed-size representation of the sentence, we perform max-pooling over the whole sentence: for each filter, we only keep the maximum score over all the positions in the sentence. We finally apply a fully connected layer to obtain a score for each topic. The binary logistic loss is used for each label independently (no softmax layer is used), since documents can have multiple labels. One neural network is used for the multi-label classification, instead of using independent ones for each label.

The Adagrad algorithm [34], with a learning rate of 0.02, is used to learn the model parameters. No normalization or zero-padding is used for convolutional neural networks. Since the dataset is imbalanced, we propose to use asymmetric costs for positive and negative examples. The ratio between these costs is denoted by the scalar α. In our experiments, H is set to 800 and α is set to 0.25 according to a grid search. The stride size of the convolutions is set to 3. We used our own Java implementation of convolutional neural network. (The implementation was validated using the "MR" dataset in [31], achieving similar performance reported in the original paper.) However, replicating our experiments with any popular deep learning framework (such as TensorFlow or Torch) should be straightforward.

Prior to training the classifiers, the following pre-processing and feature selection steps were carried out: (1) all the words in the corpus were stemmed; (2) stopwords were removed from the vocabulary; (3) dimensionality reduction were carried out by doing Named Entity Recognition (using Stanford NER [35]) to recognize Person, Location, Organization names as well as special tokens such as number, money, time. In addition, to make the comparison across tools more meaningful, we also use the word embedding input of CNN as features for SVM, examining how it differs from bag of words representations. In other words, we replace the bag-of-words input with the 100-dimensional word vectors as features, creating the system denoted as SVM-e in Table 4. For all the models we used the natural thresholds for the outputs of the classifiers, as well as for all following analyses presented in this paper.

Given limited number of samples, we are unable to split the dataset into training/validation/test with sufficient instance in each set. As such, we carried out the experiments as follows. First, we randomly split the annotated dataset into 5 subsets with roughly balanced numbers of sentences. Then for each classifier, we tuned parameters with subset 1–4 as training and subset 5 as validation, where we ran grid search and found optimal values of hyperparameters. Then we applied the hyperparameters identified on the other 4 training-validation pairs, i.e. using subset 1, 2, 3, 4 as validation set and remaining subsets as training set, respectively. Then we reported average performance on the 4 subsets.

2.5. Application to the entire community to support cross-sectional and longitudinal analyses

We applied the best-performing classifier on all sentences in the entire unannotated dataset. For each post, we assigned it topic labels that are associated with more than 1/10 of sentences in the post. As such, based on the aggregated post-level topic labels, we

Table 2
Topic labels and the number of manually annotated sentences according to each topic. For each topic, an example of manually annotated sentence is provided. The table also includes two examples with multiple labels. ALTR: alternative medicine; DAIL: daily matters; DIAG: diagnosis; FIND: findings; HSYS: health systems; MISC: miscellaneous; NUTR: nutritions; PERS: personal life; RSRC: resources; TEST: laboratory tests; TREA: treatments.

Topic	#Sentences	Example
ALTR	302	I tried everything to no avail & in desperation had acupuncture
DAIL	600	I use virgin organic coconut oil on my skin and all organic cosmetics, shampoo, conditioner, laundry detergent, household cleaner, the works!
DIAG	1127	My cancer was a 1.2 cm mucinous bc in a duct, with low growth rate
FIND	1195	I don't feel faint or anything- it just feels weird- anyone else out there had this happen?
HSYS	864	I don't know where you are located, but I would start with the Cancer Treatment Centers of America
MISC	1956	Hope this helps, cheers
NUTR	608	I am staying on a bland diet, eating every 2 h, and forcing fluids, but am worried about tomorrow based on what happened last time
PERS	1011	He has a family history of very high triglycerides
RSRC	568	I just did internet research and here is a good site with information on Curcumin
TEST	295	When I went in for my second mammogram on Dec. 18th, the radiologist told me I had to go get a biopsy based upon the mammogram
TREA	2078	I'm just curious about other warriors experience with herceptin
ALTR,NUTR	113	I read that cinnamon capsules could help with lowering glucose and ldl in our blood
HSYS,TREA	104	After dealing with the insurance company for weeks. . ...she finally started taking the Xeloda last month

are able to identify (1) what are the most prevalent topics in general in the community; (2) if there are any differences of topics among members of different cancer stages. We did not examine other factors than cancer stage in this study, because cancer stage is one particular profile information that can be accessed most comprehensively from member signatures. For each of the analysis, we take one particular factor into account: whether the post is initializing a discussion or relying to other's post. Previous studies indicate that members seek support by initializing discussions and provide support and giving feedback [6,36,8], which necessitates the distinction between initial and reply posts in our analysis. In the following parts of this paper, we use "initial post" to denote posts that initiate a thread of discussion, instead of the first posts published by members.

Armed with topic labels for each post in the dataset, we also conducted following longitudinal analyses to take timestamp into account. The primary objective for our analysis was to assess if participation in the community has an impact on topic of discussion. We thus compared distributions of topics of posts published in different periods of time with respect to user's registration date, and tracked their changes. As such, each data point is the average frequency of a topic within all posts in a given time slice (e.g., all posts published by their authors after 3 weeks of their joining the community). To show both short-term and long-term changes, three measures of time progression are used (represented as x-axis): post, day, and week.

3. Results

3.1. Manual annotation

Table 2 shows distributions and example sentences for different topics in the manually-annotated dataset. Treatment and Miscellaneous sentences are the most frequent topics in our annotated dataset, whereas Alternative Medicine and Test topics are the least prevalent. The high number of Miscellaneous sentences is explained by the fact that most posts start with greetings and end with encouragements, blessings, and signatures (all categorized as Miscellaneous in our coding).

Table 3 shows the inter-annotator agreement for each pair of annotators across the three annotators. Among the three coders, the first coder annotated all 1008 posts, while the other two complimentary coders are assigned part of the whole data set (coder 2 annotated 394, while coder 3 annotated 614). The remainder of the paper reports results on the adjudicated annotation.

3.2. Topic classification

The classifiers were evaluated using precision, recall, and F measure. In order to evaluate the overall performance of the system across all topics, micro average precision, recall and F are also calculated [37]. Micro average takes distribution of labels into consideration, and it makes more sense in this study because of the imbalance of labels in the dataset. Experiments with a baseline system are also carried out, which simply tags every sentence with all possible labels. Aggregated results for the sentence-level classification are given in Table 4 We found that CNN outperforms other models in almost all topics. In particular, CNN outperforms SVM-e which also relies on word embedding as raw features. The results suggest that both word embedding and convolutional neural network training contribute to identifying the topics.

3.3. General prevalence of topics

Prevalence of all topics at post level is given in Table 5 after applying the CNN classifier to the entire breast cancer dataset. The most prevalent topic is personal (PERS) among all posts, with 24.6% of posts labeled as such, followed by treatment (TREA, 24.6%) and diagnosis (DIAG, 9.3%). The least prevalent topics are alternative medicine (ALTR, 0.2%) and test (TEST, 1.0%). Specific to initial posts of threads, diagnosis is significantly more dominant than other topics, while popular topics among reply posts such as personal and finding are almost not found among initial posts.

In general, clinically relevant topics such as treatment, diagnosis, and finding are more prevalent than non-clinical ones, with one exception of PERS among all posts. Topic distribution in the entire breast cancer dataset is more skewed that in the annotated dataset, because the annotated dataset was sampled toward collecting more posts of rare topics such as alternative medicine (ALTR). Distribution of topics among initial posts is more uneven, suggesting that a significant amount of threads initialized by members focus on cancer diagnosis.

3.4. Topic prevalence stratified by cancer stage

In the breast cancer forum, many users self-reported disease information in their member profiles, including cancer diagnoses and treatment histories. These profile information show up in signatures when authors post, which are available to the public. One particular type of information that is mostly structured and easily extracted is cancer stage. Out of all 57,424 authors in the dataset we crawled, 17,950 (31.3%) have their cancer stage information

Table 3
Inter-rater agreements between the three topic coders measured by Cohen's Kappa. Note that coder 1 annotated all posts while coder 2 and coder 3 annotated two complimentary parts of the data. Therefore, no agreement is calculated between coder 2 and coder 3.

Label	# sen by Coder 1 and 2	Kappa	# sen by Coder 1 and 3	Kappa
Avg K		0.50		0.62
ALTR	117	0.36	185	0.29
DAIL	274	0.30	336	0.50
DIAG	301	0.50	826	0.71
FIND	370	0.56	825	0.61
HSYS	237	0.56	627	0.68
MISC	665	0.38	1291	0.76
NUTR	193	0.70	415	0.69
PERS	333	0.13	678	0.61
RSRC	159	0.63	409	0.58
TEST	100	0.69	195	0.70
TREA	690	0.67	1388	0.71

Table 4
Topic classification performance measured by F score on different topic categories, with five classifiers. bsline: the system simply tags all sentences with all 11 labels; L-LDA: the labeled LDA classifier; SVM: the SVM classifier using bag of words as features; SVM-e: the SVM classifier using word embedding as features; CNN: the convolutional neural network classifier. Bold numbers represent the best performance for individual classes.

	Baseline	L-LDA	SVM	SVM-e	CNN
Micro	19.3	54.4	57.1	59.9	**65.4**
ALTR	6.5	9.2	9.8	31.7	**35.5**
DAIL	12.5	30.1	28.6	46.9	**48.1**
DIAG	22.2	58.8	62.3	66.0	**67.1**
FIND	23.4	50.1	51.8	**61.2**	60.3
HSYS	17.5	45.4	41.0	55.4	**57.7**
MISC	35.7	76.2	76.3	72.1	**78.1**
NUTR	12.6	57.3	58.6	69.0	**72.8**
PERS	20.2	24.4	26.5	44.5	**47.8**
RSRC	11.9	48.0	48.8	55.8	**61.1**
TEST	6.3	27.6	26.0	47.8	**52.6**
TREA	37.5	65.7	68.1	66.0	**73.6**

Table 5
Percentages of all topics at post level based on automated topic classification, for all posts and initial posts respectively. Differences were measured by *t*-tests and *p*-values are reported.

All posts	**ALTR**	**DAIL**	**DIAG**	**FIND**	**HSYS**
	0.2	7.4	9.3	6.3	7.8
	NUTR	**PERS**	**RSRC**	**TEST**	**TREA**
	3.9	24.9	1.7	1.0	24.6
Initial posts	**ALTR**	**DAIL**	**DIAG**	**FIND**	**HSYS**
	0.0	0.8	46.4	1.4	7.1
	NUTR	**PERS**	**RSRC**	**TEST**	**TREA**
	0.6	8.0	2.7	0.1	22.9
p values	**ALTR**	**DAIL**	**DIAG**	**FIND**	**HSYS**
	0.54	0	0	0	0.002
	NUTR	**PERS**	**RSRC**	**TEST**	**TREA**
	0.011	0	0	0.040	0

available in signatures. Among them, 2325 are stage 0 (total number of posts: 170,610), 5968 are stage I (total number of posts: 600,500), 5907 are stage II (total number of posts: 661,990), 2447 are stage III (total number of posts: 229,955), and 2438 are stage IV (total number of posts: 460,313).

Topic distributions of posts published by members of different cancer stages are given in Fig. 2 for all posts and Fig. 3 for initial posts. For both analyses of all posts and initial posts, we did univariate *t*-tests between pairs of cancer stage groups with respect to the prevalence of certain topics. We found that most of the remarkable differences in these two figures are also statistically significant. For example, the difference between prevalence of "DIAG" among stage 0 (blue) and stage 1 (green) is not only visible in the figure, but also statistically significant according to our *t*-test. Stage 0 users focus more on cancer diagnosis and health systems, which are typical topics at early time of cancer journeys. Stage IV members, counter-intuitively, discuss more about personal lives but significantly less about treatment and clinical findings. This seems to suggest that stage IV members rely on the forum to exchange emotional more than informational support with their peers. Most differences found among all posts are even amplified among initial posts. One particular pattern among initial posts is that members with stage information, in general, post significantly less about diagnosis than other members in initial posts. One explanation might be that many of the initial posts discussing diagnosis are published by new members to the community, many of whom only posted a few times which are asking questions about whether certain signs they found indicate cancer.

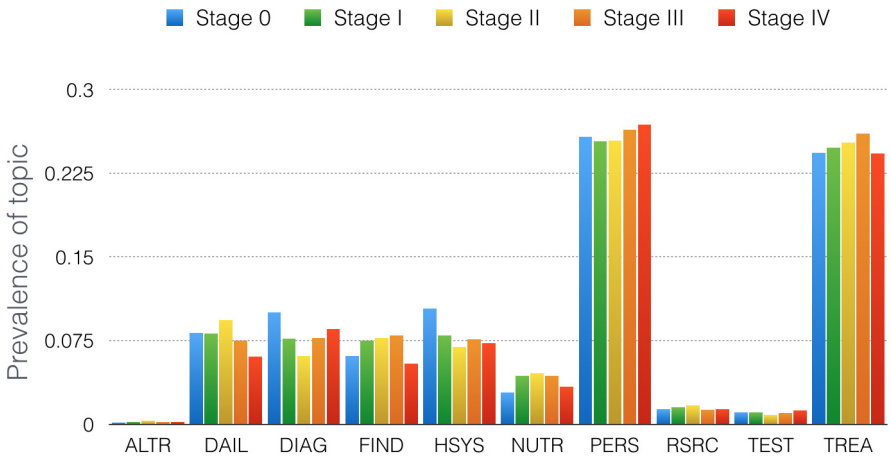

Fig. 2. Prevalence of topics of **all** posts, stratified by cancer stages of authors. *Y*-axis is the proportions of posts, among each group of patients, that are predicted by the classifier to be about the corresponding topic. For example, the blue bar over "DAIL" represents that around 7.6% of posts published by stage 0 members are about topic "daily matters". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

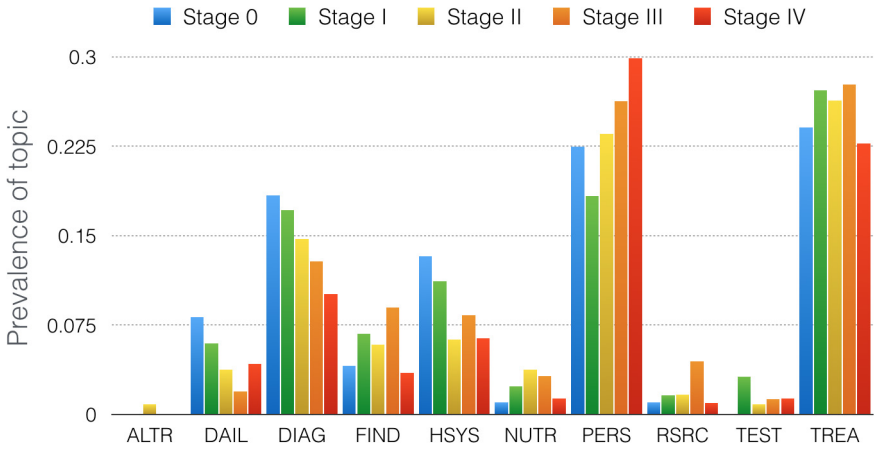

Fig. 3. Prevalence of topics of **initial** posts, stratified by cancer stages of authors. *Y*-axis is the proportions of posts, among each group of patients, that are predicted by the classifier to be about the corresponding topic. For example, the blue bar over "DAIL" represents that around 7.6% of initial posts published by stage 0 members are about topic "daily matters". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.5. Topic trajectory of users

Armed with topic labels for each post in the dataset, we conducted the following longitudinal analyses to take timestamp into account. The primary objective for our analysis was to assess if participation in the community has an impact on topic of discussion. We compared distributions of topics of posts published in different periods of time with respect to user's registration date, and tracked their changes. As such, each data point we consider is the average frequency of a topic within all posts in a given time slice (e.g., all posts published by their authors after 3 weeks of their joining the community). To visualize the changes in topic distributions through time, we plotted in addition to the individual data points fitted curves. To show both short-term and long-term changes, three measures of time progression are used (represented as *x*-axis): post, day, and week. In addition, we split our analysis by considering all posts (Fig. 4) and initial posts of discussions (Fig. 5) separately.

Several patterns are identified among all posts. First, diagnosis is the most dominant topic at early stages of participation, especially in first posts and first days. Second, prevalence of some topics such as personal (PERS), daily matters (DAIL), and nutrition (NUTR) grow steadily, while prevalence of diagnosis (DIAG) and treatment (TREA) decline as members stay longer in the community. Third, frequencies of health systems (HSYS) and findings (FIND) increase at the beginning, but slide after reaching the peaks. Finally, alternative medicine (ALTR), laboratory test (TEST), and resources (RSRC) are unpopular topics throughout members' participation. The results suggest that members' focus shifted from informational support, represented by clinically concentrated topics such as diagnosis and treatment, to emotional support, represented by personal focused on topics such as nutrition and daily lives.

Initial posts of discussions show simpler patterns. Frequency of diagnosis, as the most prevalent topic among initial posts, declines as members stay longer. Frequencies of other topics do not show clear patterns of changes.

4. Discussions

A wide range of topics are discussed in the breast cancer community, ranging from clinically relevant ones such as diagnosis and treatment to more daily matters such as nutritional supplements and stories of personal lives. In the breast cancer forum, *personal* and *treatment* are the most dominant topics, possibly representing a mix of emotional support and informational support being exchanged. When it comes to posts that initializing dis-

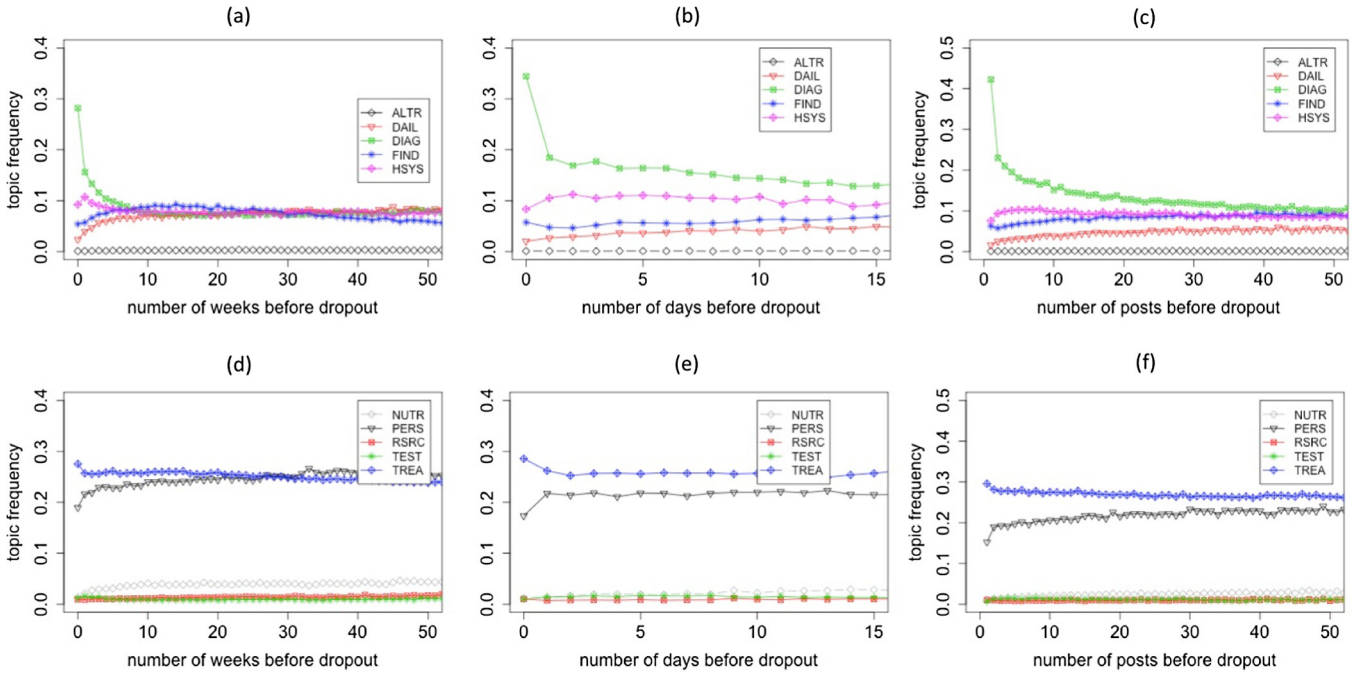

Fig. 4. How topic prevalence changes through time after members join the community. X axes represents the time point after members' first activity. Y axis is the average topic frequency of all posts that are published in the corresponding time. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

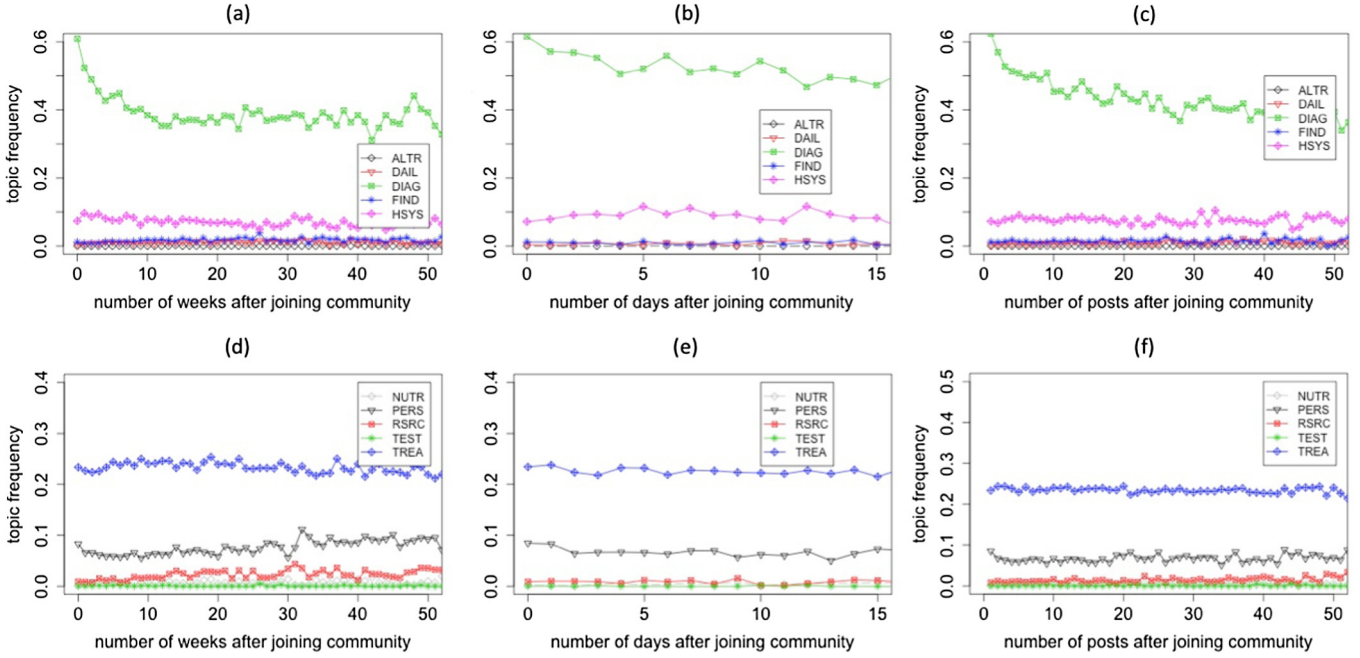

Fig. 5. How topic prevalence of initial posts of threads change through time after members join the community. X axes represents the time point after members' first activity. Y axis is the average topic frequency of all posts that are published in the corresponding time. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

cussions, cancer diagnosis is the most prevalent topic. Topics representing more personal or daily issues barely show up in initial posts, although they are quite dominant among other posts.

Cancer stage plays a role in deciding members' topics of discussions. Early stage members, many of whom are newcomers to the community, care more about diagnosis related information. Stage 0 members, in particular, focus on whether certain signs indicate cancer. They also exchange anecdotes about their experiences with healthcare providers when being diagnosed. Late stage members, such as stage IV members, usually have stayed in the community for longer time. For these members, seeking information is no longer the major motivation of participation; on the contrary, they established closer relationships with their online peers, and disclose more personal information and support with each other emotionally. It is noteworthy, however, that cancer stage information extracted from signatures may be inaccurate, since members may not report stage change in a timely manner. Also, it is naturally the case that members with late stages are more likely to

be long time users, which makes length of membership an important confounder in considering differences between members of different stages.

Finally, we found that members shifted their focus in participation, from clinically relevant topics to more casual topics as they participate longer and longer. This coincides with the difference between cancer stages, and supports that the difference is caused by length of participation more than cancer stage. Putting all the findings together, we may get a more complete picture of OHC participation with respect to topics: as members stay longer in the community, and build up closer relationship with their peers, they tend to disclose more personal information, discuss more private stories, and exchange more support emotionally; meanwhile, they seek help less but provide more, and shifted their interest from cancer diagnosis to cancer treatment.

Synthesizing the above discoveries, the difference between initial and reply posts becomes somewhat expected [38] and may be explained as follows: initial posts are more likely to be posted by new members, who ask questions and seek help more often than members who stayed in the community for longer time, while these long term members mostly provide help and reply to others' requests; meanwhile, these new members are more likely to be newly diagnosed patients focusing on cancer diagnosis, while a lot of active old members are in sessions of treatment and they exchange personal stories more often with their familiar peers. However, this cannot explain why clinical finding and health system are not prevalent among initial post, which needs to be further investigated.

One important limitation of this work is that our coding system focuses on informational classification of topics, ignoring more granular characterization of emotional and social aspects of the content. We investigate topics of OHC content from this one particularly angle, demonstrating the power of using deep neural networks to characterize the content. In future work, methods investigated in this paper could enable health researchers to discover interesting questions worth exploring about public online health communities, which can be further investigated through traditional interventional methods. Patterns identified in this work were based on quantitative and longitudinal analysis at scale, which complements traditional methods of manual content analysis, and revealed possible psycho-oncological impacts of online social support [39]. Findings of this research can be important guidance for health researchers in designing optimal interventions to deliver social support, and can also be valuable hypotheses to be examined in future clinical research.

5. Conclusion

In this paper, we provide a multi-class schema, an annotated dataset, and supervised classifiers based on convolutional neural network (CNN) and other models for the task of topic classification for online health community text. In particular, we approach the challenging multi-label multi-class topic classification task by leveraging convolutional neural networks successfully. We apply the classifier on the most popular breast cancer online community, the discussion boards of breastcancer.org, and carry out longitudinal analysis at scale to show topic distributions and topic changes throughout members' participation. Our experimental results suggest that CNN outperforms other classifiers in the task of topic classification. We also found that although personal and disease related topics are most prevalent, members of different cancer stages have different foci of topics. Finally, members change their interest as they participate, becoming increasingly interested in more personal topics in online discussions.

Conflict of interest

The authors declared that there is no conflict of interest.

Acknowledgement

This work is supported by National Institute of General Medical Sciences Grant R01GM114355.

References

[1] H. Oh, C. Rizo, M. Enkin, A. Jadad, What is eHealth (3): a systematic review of published definitions, J. Med. Internet Res. 7 (1) (2005).
[2] A. Hartzler, W. Pratt, Managing the personal side of health: how patient expertise differs from the expertise of clinicians, J. Med. Internet Res. 13 (3) (2011).
[3] K.P. Davison, J.W. Pennebaker, S.S. Dickerson, Who talks? The social psychology of illness support groups, Am. Psychol. 55 (2) (2000) 205.
[4] Y. Wang, R. Kraut, J. Levine, To stay or leave? The relationship of emotional and informational support to commitment in online health support groups, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp. 833–842.
[5] N. Elhadad, S. Zhang, P. Driscoll, S. Brody, Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions, in: AMIA Symposium.
[6] S. Zhang, E. Bantum, J. Owen, N. Elhadad, Does sustained participation in an online health community affect sentiment?, AMIA Annual Symposium Proceedings, vol 2014, American Medical Informatics Association, 2014, p. 1231.
[7] K. Portier, G.E. Greer, L. Rokach, N. Ofek, Y. Wang, P. Biyani, M. Yu, S. Banerjee, K. Zhao, P. Mitra, J. Yen.
[8] B. Qiu, K. Zhao, P. Mitra, D. Wu, C. Caragea, J. Yen, G.E. Greer, K. Portier, Get online support, feel better – sentiment analysis and dynamics in an online cancer survivor community, in: 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust, pp. 274–281.
[9] J.L. Bender, M.C. Jimenez-Marroquin, A.R. Jadad, Seeking support on Facebook: a content analysis of breast cancer groups, J. Med. Internet Res. 13 (1) (2011).
[10] P. Biyani, C. Caragea, P. Mitra, J. Yen, Identifying emotional and informational support in online health communities, in: COLING, 2014, pp. 827–836.
[11] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. (2003) 993–1022.
[12] G. Chen, J. Warren, P. Riddle, Semantic space models for classification of consumer webpages on metadata attributes, J. Biomed. Inform. 43 (5) (2010) 725–735.
[13] S. Myneni, K. Fujimoto, N. Cobb, T. Cohen, Content-driven analysis of an online community for smoking cessation: integration of qualitative techniques, automated text analysis, and affiliation networks, Am. J. Public Health 105 (6) (2015) 1206–1212.
[14] B. Sharf, Communicating breast cancer on-line: support and empowerment on the Internet, Women Health (February 2014) (1997) 37–41.
[15] A.C. Lewallen, J.E. Owen, E.O. Bantum, A.L. Stanton, How language affects peer responsiveness in an online cancer support group: implications for treatment design and facilitation, Psycho-oncology 23 (7) (2014) 766–772.
[16] J.E. Owen, J.C. Klapow, D.L. Roth, D.C. Tucker, Use of the internet for information and support: disclosure among persons with breast and prostate cancer, J. Behav. Med. 27 (5) (2004) 491–505.
[17] A. Civan, W. Pratt, Threading together patient expertise, AMIA Annual Symposium Proceedings, vol. 2007, American Medical Informatics Association, 2007, p. 140.
[18] A. Meier, E.J. Lyons, G. Frydman, M. Forlenza, B.K. Rimer, How cancer survivors provide support on cancer-related Internet mailing lists, J. Med. Internet Res. 9 (2) (2007) e12.
[19] M. Cappiello, R.S. Cunningham, M.T. Knobf, D. Erdos, Breast cancer survivors: information and support after treatment, Clin. Nurs. Res. 16 (4) (2007) 278–293 (discussion 294–301).
[20] E. Loper, S. Bird, NLTK: the natural language toolkit, Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Association for Computational Linguistics, vol. 1, 2002, pp. 63–70.
[21] T.O. Blank, S.D. Schmidt, S.A. Vangsness, A.K. Monteiro, P.V. Santagata, Differences among breast and prostate cancer online support groups, Comput. Human Behav. 26 (6) (2010) 1400–1404.
[22] M.M. Skeels, K.T. Unruh, C. Powell, W. Pratt, Catalyzing social support for breast cancer patients, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2010, pp. 173–182.
[23] K.-Y. Wen, F. McTavish, G. Kreps, M. Wise, D. Gustafson, From diagnosis to death: a case study of coping with breast cancer as seen through online discussion group messages, J. Comput.-Mediated Commun. 16 (2) (2011) 331–361.
[24] J.L. Bender, J. Katz, L.E. Ferris, A.R. Jadad, What is the role of online support from the perspective of facilitators of face-to-face support groups? A multi-method study of the use of breast cancer online communities, Patient Educ. Couns. 93 (3) (2013) 472–479.

[25] S.C. Kim, D.V. Shah, K. Namkoong, F.M. McTavish, D.H. Gustafson, Predictors of online health information seeking among women with breast cancer: the role of social support perception and emotional well-being, J. Comput.-Mediated Commun. 18 (2) (2013) 98–118.
[26] S. Bird, M. Liberman, A formal framework for linguistic annotation, Speech Commun. 33 (1) (2001) 23–60.
[27] J. Cohen et al., A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46.
[28] G. Tsoumakas, I. Katakis, Multi-label classification: an overview, Int. J. Data Warehousing Min. (IJDWM) 3 (3) (2007) 1–13.
[29] D. Ramage, D. Hall, R. Nallapati, C.D. Manning, Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora.
[30] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (3) (1999) 293–300.
[31] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751.
[32] G. Heinrich, Parameter Estimation for Text Analysis, University of Leipzig, Tech. Rep.
[33] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011). 27:1–27:27.
[34] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (2011) 2121–2159.
[35] J.R. Finkel, T. Grenager, C. Manning, Incorporating non-local information into information extraction systems by gibbs sampling, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2005, pp. 363–370.
[36] E. Kim, J. Han, T. Moon, B. Shaw, The process and effect of supportive message expression and reception in online breast cancer support groups, Psycho-oncology 21 (5) (2012) 531–540, http://dx.doi.org/10.1002/pon.1942.
[37] Y. Yang, An evaluation of statistical approaches to text categorization, Inform. Retriev. 1 (1–2) (1999) 69–90.
[38] B. Qiu, K. Zhao, P. Mitra, D. Wu, C. Caragea, J. Yen, G.E. Greer, K. Portier, Get online support, feel better–sentiment analysis and dynamics in an online cancer survivor community, in: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on, 2011 IEEE Third International Conference on Social Computing (SocialCom), IEEE, 2011, pp. 274–281.
[39] R. McArthur, P. Bruza, J. Warren, D. Kralik, Projecting computational sense of self: a study of transition in a chronic illness online community, Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), vol. 5, IEEE, 2006. 91c–91c.